Dr. Straw's Tips for Success in D209

Please send all questions and suggestions with the subject: "D209 tips suggestion" to eric.straw@wgu.edu.

These tips provide my suggestions as well as answers to the most common student questions. Tips are organized alphabetically within each task requirements section (e.g. A, B, C, etc.) and within the *General* section, which comes before the task requirements sections.

General

D209 vs D208

D209 has both similarities and differences to D208. You can re-use your work from D208 in D209 where the two courses are similar.

Similar: Selection of outcome variable. D209 classification (Task 1) requires a categorical outcome variable, which means you can use the dependent variable from D208 logistic regression (Task 2). D209 prediction (Task 2) requires a continuous outcome variable, which means you can use the dependent variable from D208 linear regression (Task 1).

Different: D209 requires a different type of research question. Both tasks in D209 are attempting to estimate (i.e. determine or predict) the value of the outcome variable. Neither classification (Task 1) nor prediction (Task 2) are used to find causation (i.e. Does B cause A). Rather these methods are used to estimate the outcome (i.e. Can B predict A).

Different: D209 requires some different data preparation steps, such as splitting your data 70% training and 30% test, using a different rule for dummy variable creation, normalizing continuous variables for KNN, etc.

Different: D209 requires running only one model (i.e. there is no model reduction built into the tasks). This means you are not required to do any hyperparameter tuning.

Different: The model you choose in D209 will determine the performance metrics (requirement E1) you should use, and most performance metrics are similar to those used in D208.

Data and Data Dictionary

Do not use the data from D206. Ensure you have downloaded the data and data dictionary for D207/D208/D209.

1. Go to the D209 course page
2. Select View Task under Assessments at the bottom center of the page
3. Scroll to the bottom of the page and select the Data Sets and Associated Data Dictionaries link
4. Select the link for the data set you will be using
5. Unzip the downloaded folder
6. The data file is in CSV format
7. The data dictionary is in PDF format. Ignore the Scenario on page 1 of the PDF. The Scenario has nothing to do with your work in this class.

Dr. Straw's Tips for Success in D209

Data Dictionary: PDF Scenarios

The scenarios described on page 1 of both the churn dataset and the medical dataset are just examples. Even though these say, "You have been asked to..." it does not mean that you should use the example scenario as the basis for your analysis in D209. In fact, both example scenarios use a categorical outcome variable, which is appropriate for classification (Task 1) but not for prediction (Task 2) because prediction requires a continuous outcome variable. Thus, you will need to select a continuous variable for your outcome variable for prediction (Task 2).

DataCamp: Data Files

Do the following to access the data files for the resources in DataCamp.

(1) From the custom track in DataCamp (i.e. the landing page), select a course title.

(2) You will find the data files for that course at the bottom right corner of the page.

Python data files are in CSV format. R data files are in FST (fast storage) format. These FST files require the fst package and use of read_fst().

DataCamp: PDF of Slides

You can download a PDF of the slides for a DataCamp chapter by selecting the page icon in the upper right corner of any of the chapter's videos. Having these slides available will make your studies more efficient because you will not need to search online for syntax help as you complete the demonstration portion after each video.

You can also view the slides on the Slides tab next to the Console in the exercises. However, this view is quite small and challenging to use.

Task 1 vs Task 2

Task 1 is classification with KNN or naive Bayes. Classification is used when the outcome variable is categorical. Thus, Task 1 requires a categorical outcome variable.

Task 2 is prediction with regression decision tree, regression random forest, lasso regression, or ridge regression. Regression models are used when the outcome variable is continuous. Thus, Task 2 requires a continuous outcome variable and requires that you use regression examples to guide you.

Textbooks

Data Science Using Python and R, which was the textbook in D206, provides good supplemental content for this course. Relevant material includes the following chapters. You can find additional helpful content by searching this book.

Chapter 6 Decision Trees & Random Forests
Chapter 7 Model Evaluation
Chapter 8 Naive Bayes

Practical Statistics for Data Scientists is also a great supplemental resource in this course. Relevant material includes the following chapters. You can find additional helpful content by searching this book.

Chapter 4: Regression and Prediction
Chapter 5: Classification
Chapter 6: Statistical Machine Learning

Dr. Straw's Tips for Success in D209

Section A

Research Questions

Both tasks in D209 are attempting to estimate (i.e. determine or predict) the value of the outcome variable (e.g. Can B predict A). Thus, your research question should reflect this goal.

Section B

Model Assumptions

Here is a good article that explores the assumptions of eight different model building techniques -- https://medium.com/swlh/its-all-about-assumptions-pros-cons-497783cfed2d

Section C

Dummy Variables

For machine learning algorithms such as KNN (D209 Task 1) and regression trees (D209 Task 2), we use the same number of dummy variables as categories in our categorical variable.

We create dummy variables (also called indicator variables) to represent categorical variables. Dummy variables always contain values of 1 or 0. For example, a categorical variable may contain one of three values (i.e. k=3): Yes, No, or Maybe. For KNN (D209 Task 1) and regression trees (D209 Task 2), we would use three dummy variables: Yes, No, and Maybe. Each of these new variables would have possible values of 0 or 1.

This rule is different than the k-1 rule used in D208, which was based on the need to prevent multicollinearity in linear regression and logistic regression. See Shmueli (2015) for more details at http://www.bzst.com/2015/08/categorical-predictors-how-many-dummies.html

C2: Classifying variable types

C2 requires you to "classify each variable as continuous or categorical". This is improper usage of the term "continuous". You should classify all variables as either numeric or categorical.

Numeric variables can be either discrete (e.g. number of children) or continuous (e.g. weight).

For more information on variable types see my breakdown of data types PDF file in the D209 Student Resource folder. The link to this folder can be located by selecting Course Tips on the right-hand side of the course page, then selecting View All.

Dr. Straw's Tips for Success in D209

Section D

D2: Intermediate Calculations

Section D2 asks you to provide "screenshots of the intermediate calculations performed". You can interpret the phrase "intermediate calculations" to include any calculations you performed to enable you to run your model or any calculations you performed to measure the performance of your model.

Please note that using a function, e.g. something like mean_squared_error(Y_true,y_predict), is not performing a calculation. A calculation involves coding math into a formula to arrive at an answer.

Include a statement like one of the following statements in D2: (a) "I have included screenshots of all the intermediate calculations I performed" or (b) "I did not perform any intermediate calculations."

Hyperparameter Tuning

You are not required to perform hyperparameter tuning of your models in D209.

You can learn a lot about modeling by attempting to improve your models (a.k.a hyperparameter tuning). Jordan (2017) has a good article discussing hyperparameter tuning methods.

Jordan, J. (2017). Hyperparameter tuning for machine learning models. Available at https://www.jeremyjordan.me/hyperparameter-tuning/

Splitting the Data

Use a 70/30 ratio for splitting your data into a Training set and a Test set. Use a method to randomly select 70% of your data and copy this data to a Training set. Copy the remaining 30% of the data to a Test set. Build your models with the Training data set. Evaluate your models with the Testing data set.

Using a 70/30 split will satisfy the D1 requirement in both tasks. Using less than 70% for the Training data set will result in your submission being returned for revisions.

Section E

Accuracy in Task 2: Requirement E1

D209 Task 2 is for prediction with a regression model, which is used when the outcome variable is continuous. Requirement E1 states, "Explain the accuracy...". You should interpret this requirement as "accuracy" rather than "Accuracy". There is no "Accuracy" measurement for the prediction methods required in Task 2 (i.e. there is no formula that provides a specific number that is labeled "Accuracy"). Rather, you can evaluate the accuracy of your prediction with model evaluation metrics such as MSE, RMSE, and R-Squared to meet the E1 requirement.

Confusion Matrix in R

You will receive an error on the R confusionMatrix() function if the variables are not factors and if the variables have different levels. A factor is the name for a categorical variable in R. You can use either the as.factor() or factor() function to ensure your variables are treated as factors. You must also ensure your variables have the same levels. Factor variables have categories (e.g. hot or cold; 1, 2, or 3; etc.). In the language of R, these are called levels. Thus, a factor variable has levels. You can use either the factor() or levels() function to ensure your variables in the confusionMatrix() function have the same number of levels.

Here is an example confusionMatrix() statement:

```
confusionMatrix(as.factor(MyLogisticModel$MyDependentVariable),
as.factor(MyPredictedVariable))
```

Mean Squared Error (MSE) in Task 2: Requirement E1

D209 Task 2 requirement E1 states, "Explain the accuracy and the mean squared error (MSE) of your prediction model."

MSE is a model evaluation metric for models with a continuous target variable. Thus, you should use a continuous variable as your target variable in Task 2 because you cannot provide the MSE if you have a categorical target variable.

However, students have successfully passed D209 Task 2 using a categorical target variable by doing the following:

(1) Explaining why MSE is not an appropriate model evaluation metric for the chosen target variable.

(2) Identify and provide the appropriate model evaluation metrics for the chosen target variable.

I recommend that you avoid this tactic.

Section F

Panopto

Your Panopto videos are one way you will demonstrate your solution to each task. You must narrate and explain your code in your Panopto videos in D209, and you will need to create a Panopto video for each task. You must be visible in the Panopto videos, and your computer screen must be visible in the Panopto videos. In addition, you must explain your programming environment including type and version of operating system, integrated development environment (IDE), and programming language.

There are three links at the bottom of the task overview page: (1) Panopto Access; (2) Panopto FAQs; and (3) Panopto how-to videos. I encourage you to ensure you have Panopto access as soon as possible. This access allows you to place your completed video in the D206 course folder. You must still upload your Panopto video with your task submission.

Section H

References

You are not required to follow APA or any other strict writing guide for references and citations. However, APA provides an adequate format to emulate.

Every item listed in section H must be cited in your paper. For suggestions on how to write in-text citations see the first two links in the Create In-Text Citations section at https://cm.wgu.edu/t5/Writing-Center-Knowledge-Base/I-Need-Help-with-APA-Style/ta-p/33524.

A few details:

- Use the last names of authors (e.g. LoDolce in the Format References: Basic Principles example via the link above) or, if the last names are not provided, use the publisher name (e.g. Obesity Action Coalition in the Format References: Basic Principles example via the link above).
- n.d. means No Date. Use either the date of the publication or use n.d. if the date is not provided by the publisher.